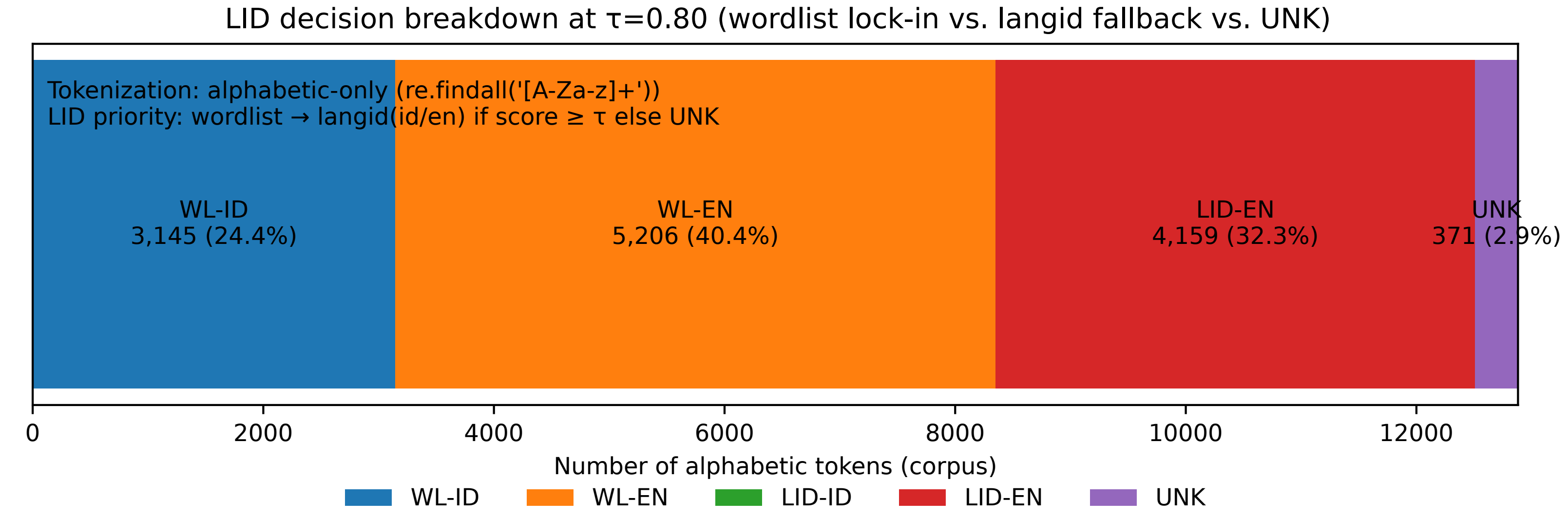LID decision breakdown at τ=0.80 (wordlist lock-in vs. langid fallback vs. UNK)
Tokenization: alphabetic-only (re.findall('[A-Za-z]+'))
LID priority: wordlist → langid(id/en) if score ≥ τ else UNK
WL-ID
3,145 (24.4%)
WL-EN
5,206 (40.4%)
LID-EN
4,159 (32.3%)
UNK
371 (2.9%)
0
2000
4000
6000
8000
10000
12000
Number of alphabetic tokens (corpus)
WL-ID
WL-EN
LID-ID
LID-EN
UNK